Jon Braswell
Finite Mixture Homework

In-Class Assignment 1

Problem Statement:

Write a procedure to generate r.v. from normal kernel density estimation.

Solution:

Use each data point as the center of a normal distribution, with your h value as the variance. Take N random samples from the data with replacement. Then use each of your sample points as the center of a normal distribution with a variance of H. Sample from each of these distributions to create your data.

```
N = 1000;
xs = zeros(1,N);
idx = datasample(data,N);
xs = normrnd(idx,hn);
hist(xs)
```

In-Class Assignment 2

Problem Statement

Create artificial 3-term mixture data specified in the in class notes. Find the finite mixture model using EM algorithm. Use starting parameters listed in notes. Generate 1500 random variables from your FM model. Draw histogram of raw data and data generated to compare.

Jon Braswell
Finite Mixture Homework

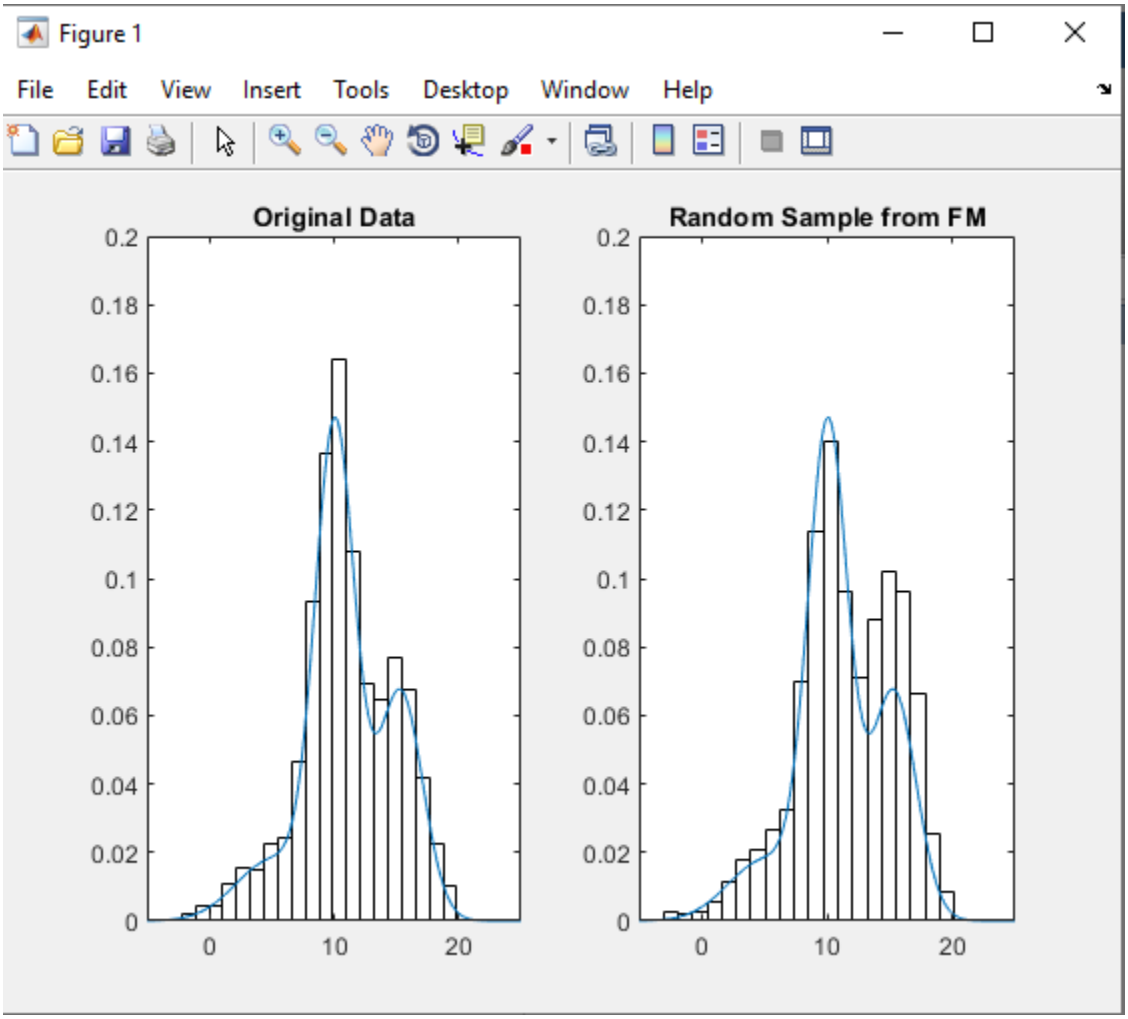
Code:

```
% 2a
n=1500;
nx = [200,800,500];
mu=[5 10 15];
sigma=[3 1.5 2];
data = zeros(n,1);
data(1:nx(1)) = normrnd(mu(1),sigma(1),nx(1),1);
data(nx(1)+1:nx(1)+nx(2)) = normrnd(mu(2),sigma(2),nx(2),1);
data(nx(1)+nx(2)+1:n) = normrnd(mu(3),sigma(3),nx(3),1);
% 2b
muin = [4 12 16];
piesin = [.333 .333 .333];
varin = [2 1 2];
max_it = 100;
tol = 0.0001;
[pies,mus,vars]=...
    csfinmix(data,muin,varin,piesin,max_it,tol);
xx=-5:.1:25;
fhat = zeros(size(xx));
for i=1:3
   fhat = fhat+pies(i)*normpdf(xx,mus(i),sqrt(vars(i)));
end
plot(xx,fhat)
% 2c
N=1500;
x = zeros(N,1);
r = rand(N,1);
ind1 = length(find(r <= pies(1)));
ind2 = length(find(r <= pies(2)));

x(1:ind1) = normrnd(mus(1),sqrt(vars(1)),ind1,1);
x(ind1+1:ind2) = normrnd(mus(2),sqrt(vars(2)),ind2-ind1,1);
x(ind2+1:N) = normrnd(mus(3),sqrt(vars(3)),N-ind2,1);
% 2d
figure(1)
subplot(121)
[cnt,data]=hist(data,20);
bar(data,cnt/n,1,'w')
title('Original Data')
axis([-5 25 0 .2])
hold on
plot(xx,fhat)
hold off
subplot(122)
[cnt,x]=hist(x,20);
```

Jon Braswell
Finite Mixture Homework

Results



Discussion

The histogram of the random sample looks very similar to that of the initial data. Imposing the theoretical underlying distribution over both confirms their similarity.

In Class Assignment 3

Problem Statement:

Repeat Problem 2, but use Kernel Density Method instead.

Jon Braswell
Finite Mixture Homework

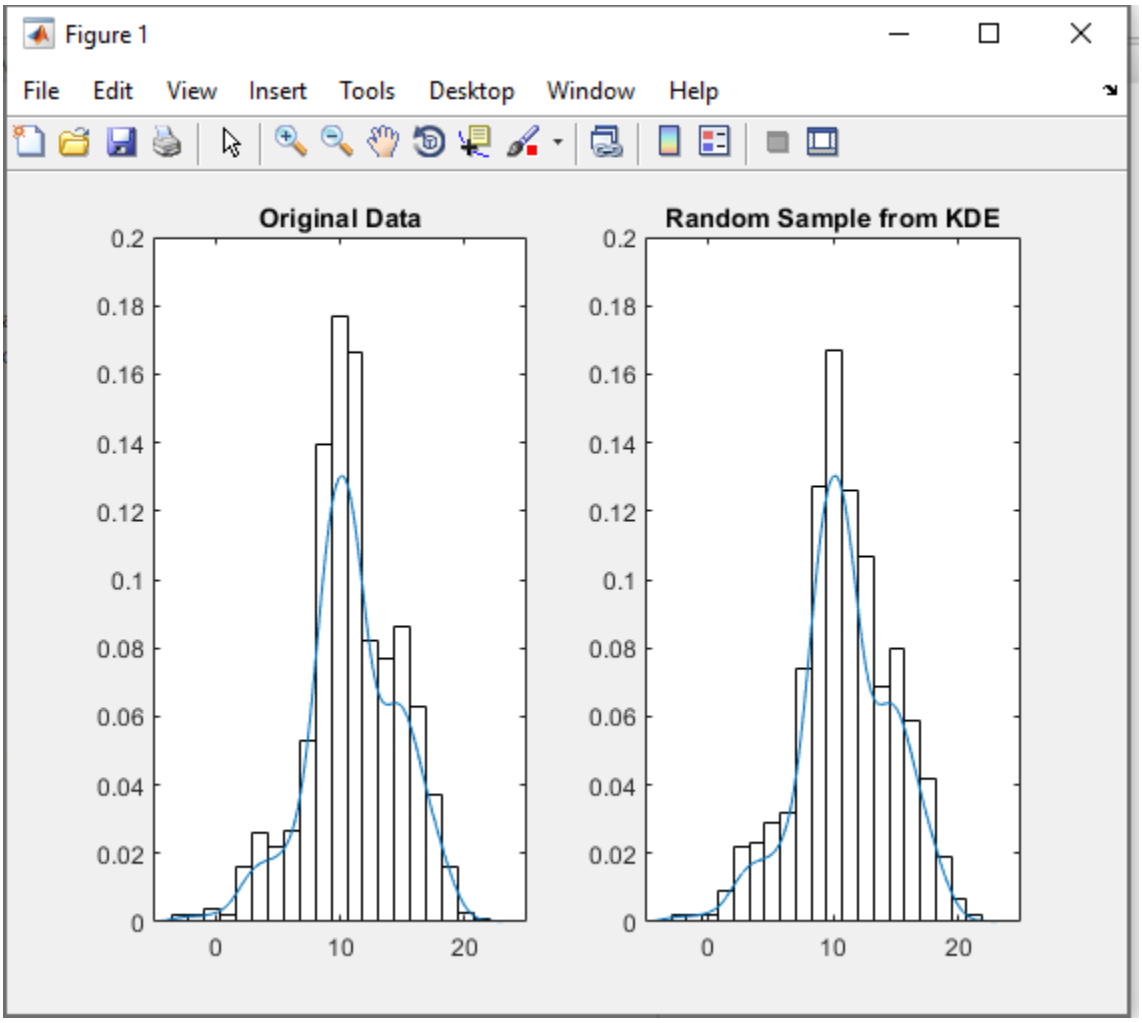
Code:

```
n=1500;
nx = [200,800,500];
mu=[5 10 15];
sigma=[3 1.5 2];
data = zeros(n,1);

data(1:nx(1)) = normrnd(mu(1),sigma(1),nx(1),1);
data(nx(1)+1:nx(1)+nx(2)) = normrnd(mu(2),sigma(2),nx(2),1);
data(nx(1)+nx(2)+1:n) = normrnd(mu(3),sigma(3),nx(3),1);
% 3B
n = length(data);
t0 = min(data)-1;
tm = max(data)+1;
len = 1500;
x = linspace(t0,tm,len);
fhatN = zeros(size(x));
hn = 1.06*n^(-1/5)*std(data);
for i=1:n
    f=exp(-(1/(2*hn^2))*(x-data(i)).^2)/sqrt(2*pi)/hn;
    fhatN = fhatN+f/(n);
end
% 3C
N = 1000;
xs = zeros(1,N);
idx = datasample(data,N);
xs = normrnd(idx,hn);
hist(xs)
% 3D
figure(1)
subplot(121)
[cnt,data]=hist(data,20);
bar(data,cnt/n,1,'w')
title('Original Data')
axis([-5 25 0 .2])
hold on
plot(x,fhatN)
hold off
subplot(122)
[cnt,xs]=hist(xs,20);
bar(xs,cnt/N,1,'w')
title('Random Sample from KDE')
axis([-5 25 0 .2])
hold on
plot(x,fhatN)
hold off
```

Jon Braswell
Finite Mixture Homework

Results



Discussion

The histogram of the random sample looks very similar to that of the initial data. Imposing the theoretical underlying distribution over both confirms their similarity. A rough glance suggests that the Finite Mixture method does a better job of capturing the distribution.

Jon Braswell
Finite Mixture Homework

Textbook Problem #3

Problem Statement

Generate 100 normal random variables and construct a histogram. Calculate the MSE and MAE Monte Carlo simulation. Do this for varying bin widths, what is the better width? Does the sample size make a difference? Does using points at the center vs the edge make a difference?

Code

```
sims = 100;
n = 100;
h = [.1,.5,1,1.5,2];
t0   = -4;
tm   = 4;
MSEs = zeros(1,length(h));
MAEs = zeros(1,length(h));
x0 = 1;
for j = 1:length(h)
    rng  = tm-t0;
    nbin = ceil(rng/h(j));
    bins = t0:h(j):(nbin*h(j)+t0);

    [k,dist] = dsearchn(bins',x0);
    errors = zeros(1,sims);

    for i = 1:sims
        x = randn(1,n);
        counts = histc(x,bins);
        fhat = counts/sum(counts);
        errors(i) = fhat(k) - normpdf(x0);
    end
    hist(x,bins)
    MSEs(j) = mean(errors.^2);
    MAEs(j) = mean(abs(errors));
end
MSEs
MAEs
T1 = table(h');
T2 = table(MSEs');
T3 = table(MAEs');
T = table(h',MSEs',MAEs');
T.Properties.VariableNames = {'Bin_Width','MSE','MAE'};
T
```

Jon Braswell
Finite Mixture Homework

Output

Using X = 0

Bin_Width	MSE	MAE
0.1	0.13018	0.36024
0.5	0.04462	0.20754
1	0.0049919	0.060259
1.5	0.016353	0.12016
2	0.0088907	0.079231

Using X = 1

Bin_Width	MSE	MAE
0.1	0.048118	0.21887
0.5	0.023676	0.15157
1	0.012621	0.10696
1.5	0.0035096	0.047636
2	0.055028	0.22973

Using X = 2

Bin_Width	MSE	MAE
0.1	0.0023237	0.047691
0.5	0.0015691	0.037091
1	0.0012475	0.032569
1.5	0.0012184	0.031731
2	0.0011919	0.03141

Discussion

For all cases, bin widths between .5 and 1.5 resulted in better density estimations. It was also found that the MSE and MAE were lower for points in the extremity than points in the center of the distribution.

Jon Braswell
Finite Mixture Homework

Textbook Problem #22

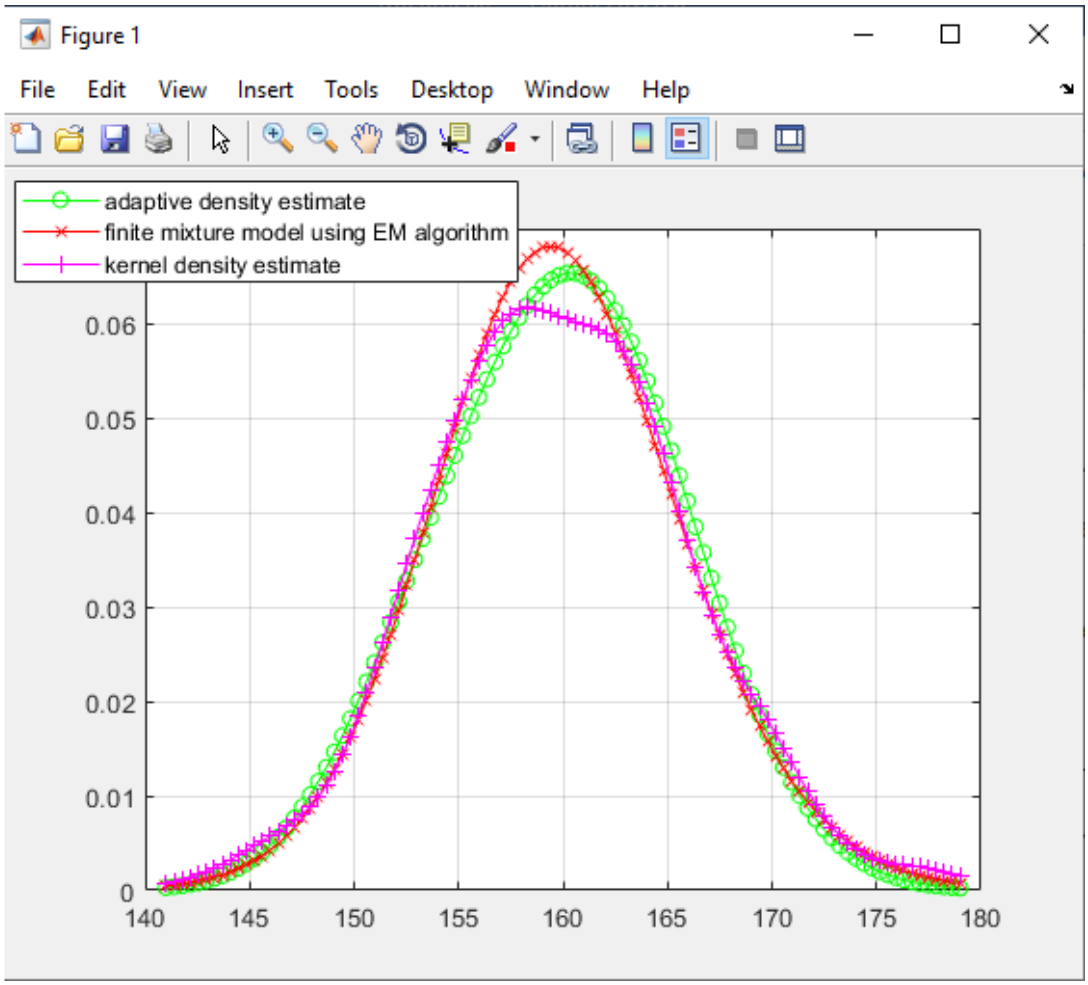
Use some of the univariate density estimation techniques from this chapter to explore the elderly data. Is there any evidence of multiple distributions?

Code

```
clear all;
close all;
clc;
addpath('../../Computational Stats/Toolbox')
load elderly;
data = heights;
% Adaptive Density Estimate. Based on histogram guess at 2 distinct groups.
maxterms = 2;
[pihat_ml,muhat_ml,varhat_ml] = csadpmix(data,maxterms);
xinterp = linspace( min(data)-1.0, max(data)+1.0 );
fhat_1 = zeros(1,length(xinterp));
for ii=1:length(pihat_ml),
  fhat_1 = fhat_1 + pihat_ml(ii)*normpdf(xinterp,muhat_ml(ii),sqrt(varhat_ml(ii)
end
fh=figure; ph1=plot( xinterp, fhat_1, '-og' ); hold on; grid on;
% Finite Mixture With EM Algo
max_its = 100000; tol = 1e-9;
[pi_ml_em,mu_ml_em,var_ml_em] = csfinmix(data,muhat_ml,varhat_ml,pihat_ml,max_it
xinterp = linspace( min(data)-1.0, max(data)+1.0 );
fhat_1 = zeros(1,length(xinterp));
for ii=1:length(pi_ml_em),
  fhat_1 = fhat_1 + pi_ml_em(ii)*normpdf(xinterp,mu_ml_em(ii),sqrt(var_ml_em(ii)
end
figure(fh); ph2=plot( xinterp, fhat_1, '-xr' ); hold on; grid on;
%Kernel Desnity Estimation
n = length(data);
t0 = min(data)-1;
tm = max(data)+1;
len = 1000;
xinterp = linspace( min(data)-1.0, max(data)+1.0 );
fhatN = zeros(size(xinterp));
hn = 1.06*n^(-1/5)*std(data);
for i=1:n
   f=exp(-(1/(2*hn^2))*(xinterp-data(i)).^2)/sqrt(2*pi)/hn;
   fhatN = fhatN+f/(n);
end
figure(fh); ph3=plot( xinterp, fhatN, '-+m' ); hold on; grid on;
legend( [ph1,ph2,ph3], {'adaptive density estimate','finite mixture model using
```

Jon Braswell
Finite Mixture Homework

Results



Discussion

The approach to the problem used the Adaptive Density Estimate found through the algorithm provided by the textbook. Looking at the histogram of the data it was determined that there were at most 2 distributions from which this data was coming, evidenced by a weak bump in the right tail. The adaptive density estimation provided some initial values for means, variances and priors. These were then fed into a finite mixture model utilizing the EM algorithm. The resulting distribution was plotted. Finally, a Kernel Density Estimate was done. All three were plotted above. The results suggest that the underlying distribution is normal, and there is no evidence of multiple distributions underlying it,